MODELLING, SIMULATION, AND ANALYSIS OF SEQUENCE-BASED MODELS FOR SMART LIGHTING VOICE COMMAND CLASSIFIERS WITH MFCC-BASED DATA AUGMENTATION

Yohanes Batara Setya[1], Feddy Setio Pribadi[2*]
Department of Electrical Engineering[1,2], Faculty of Engineering, Universitas Negeri Semarang
Building E11, 2st Floor, Sekaran, Gunungpati, Semarang, Central Java, 50229, Indonesia
*email : yohanesbatarasetya@students.unnes.ac.id , feddy.setio@mail.unnes.ac.id **

Abstract - *Voice command classification is essential for smart lighting systems in IoT environments. However, existing approaches often struggle in real-world scenarios with background noise and speaker variability due to limited and imbalanced training data. This indicates a need for models that maintain high accuracy under such conditions. To address this, the study evaluates three deep learning architectures: a Deep Neural Network (DNN), a Gated Recurrent Unit (GRU), and a bidirectional Long Short-Term Memory (LSTM) network, run on the Google Speech Commands dataset. The classification targets six voice commands ("right", "off", "left", "on", "down", "up") using Mel-Frequency Cepstral Coefficients (MFCCs) as features. Data augmentation techniques, including pitch shifting, time stretching, mix-up, and noise injection, are used to expand the dataset, balance class distributions, and simulate acoustic conditions such as background noise and speaker differences. Model performance is assessed through confusion matrices and receiver operating characteristic curves (ROC-AUC) across training, validation, and test sets. The bidirectional LSTM achieves the highest test accuracy (94%), followed by GRU (92%) and DNN (79%). The LSTM model also generalizes well, showing no signs of overfitting and maintaining stable performance in the presence of acoustic variation. These results suggest that combining bidirectional LSTM with MFCC-based augmentation provides a more robust approach to voice command recognition, particularly in IoT-based smart lighting contexts, where environmental variability is common.*

Keywords - *Voice Command Classification, Smart Lighting, Data Augmentation, Bi-LSTM, DNN, GRU, MFCC Features, Temporal Dependencies.*

I. INTRODUCTION

The development of Internet of Things (IoT) devices has revolutionized smart environments by enabling seamless human-computer interaction through voice-controlled systems [1],[2]. Voice command classification is particularly valuable in smart lighting applications, offering efficient control over lighting conditions, energy management, and accessibility. However, the deployment of these systems faces notable challenges, including background noise, speaker variability, and limited annotated training data, which can significantly impair model robustness and reliability in real-world acoustic conditions [3]. Conventional machine learning methods relying on handcrafted features and shallow models often lack the flexibility to generalize effectively across diverse environments [4]. To address these limitations, recent research has demonstrated the efficacy of deep learning architectures, particularly LSTM networks, in capturing temporal and contextual dependencies in sequential audio data [5],[6],[7]. LSTMs have shown considerable promise in tasks such as voice activity detection, automatic speech recognition (ASR), and speaker identification. Nonetheless, IoT-specific applications remain challenged by small datasets and susceptibility to overfitting. Data augmentation strategies, including pitch shifting, time stretching, and noise injection, have been employed to synthetically enlarge training data and improve model resilience [8],[9]. Huh et al. [3] highlighted the benefits of SpecAugment and Gaussian noise for improving

phoneme and ASR models, while Alharbi et al. [4] emphasized the critical role of augmentation in enhancing deep learning-based ASR systems under noisy conditions.

Building on these advancements, this study proposes a novel framework for voice command recognition tailored to IoT-based smart lighting systems. The proposed approach MFCC features with a Bi-LSTM architecture and advanced data augmentation techniques. Previous works, such as Alex et al. [8] demonstrated the importance of augmentation strategies in improving model robustness for speech separation and accent-aware ASR, respectively. Additionally, practical implementations in IoT systems by Biswal et al. [10] have showcased the operational feasibility of voice-controlled lighting applications. The key contributions of this study are: (1) the design of a Bi-LSTM model optimized for IoT lighting control, (2) a systematic assessment of augmentation techniques to enhance classification robustness, and (3) empirical validation demonstrating improved performance and resilience to acoustic variability.

Considering these challenges like particularly background noise, speaker variability, and limited annotated training data, there remains a critical question: Can a Bi-LSTM architecture combined with advanced MFCC-based data augmentation significantly improve the robustness and accuracy of voice command classification in IoT-based smart lighting systems? This research answers this question by systematically evaluating whether the proposed method outperforms conventional deep learning models (DNN and GRU) under realistic acoustic conditions. Specifically, we hypothesize that integrating bidirectional temporal modeling and diverse data augmentation strategies will enhance model resilience, resulting in superior classification performance and reduced susceptibility to overfitting.

II. SIGNIFICANCE OF THE STUDY

A. Literature Review

While prior studies have established the effectiveness of sequential deep learning models such as LSTM and GRU, along with augmentation strategies like pitch shifting and SpecAugment, these techniques have rarely been adapted specifically to the domain of smart lighting within IoT ecosystems [11],[12]. Most previous research has focused on general-purpose automatic speech recognition or speech separation, often using large-scale datasets and overlooking critical constraints inherent in IoT environments, such as limited processing power, ambient noise, and variability among speakers. Although implementations like that of Biswal et al. [10] have demonstrated the feasibility of voice-activated lighting systems, they lack systematic evaluations of architectural robustness and performance in noisy, real-world conditions.

This study fills that gap by proposing and validating a tailored voice command classification framework for smart lighting using a Bi-LSTM model enhanced with MFCC-based feature extraction and advanced data augmentation. Abdul and Al-Talabani [13] reviewed MFCC methodologies and applications, highlighting their efficiency and reliability in capturing phonetic features, which supports the use of MFCC in this framework. The approach not only improves recognition accuracy but also mitigates overfitting, adapts well to user variability, and performs reliably in acoustically diverse scenarios. The research further expands the system's functionality to include speaker identification, enhancing personalization and security in smart environments. By grounding the method in real-world IoT constraints and rigorously benchmarking it against DNN and GRU baselines, this work delivers a domain-specific, empirically validated solution. It bridges the disconnect between generic voice recognition research and the practical demands of smart lighting applications, offering a robust and efficient framework that can serve as a reference for future development in intelligent voice-controlled systems.

B. Methodology

The methodology for the voice command recognition study begins with Exploratory Data Analysis (EDA) on the balanced Google Speech Commands Dataset, which includes six classes: "right", "off", "left", "on", "down", and "up". This balance helps prevent model bias. We perform Feature Extraction using MFCC to derive meaningful features from raw audio, followed by Data Augmentation to create audio variations for better model adaptability [14]. After preparing the dataset by splitting it into training, validation, and test sets, we design neural network architectures such as DNN, GRU, and LSTM, optimized with an Adam optimizer and ExponentialDecay learning rate. During model training, we use early stopping to avoid overfitting, concluding with model validation and testing, where results are visualized via loss curves and confusion matrices.

To further enhance our study, we adapt our approach to identify voice characteristics, distinguishing "User 1", "User 2", and "Unknown User". Due to the class imbalance, especially with "Unknown User" samples, we apply advanced data augmentation for minority classes and layer normalization within the Bi-LSTM model to improve robustness and handle intra-class variability in speaking styles effectively.

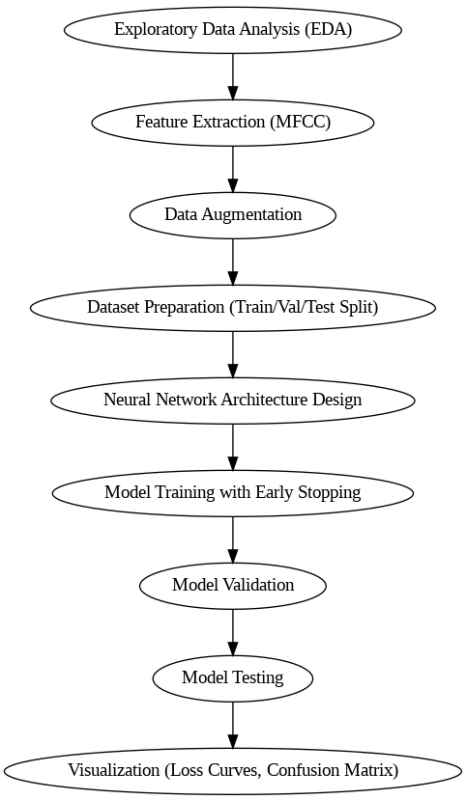

Figure 1. Simulation Flowchart

Figure 1 showing the flowchart that shows this narrative flow as a visual roadmap. Starting with an EDA oval, which reflects our initial progress into the balanced Google Speech Command Dataset, leading to Feature Extraction for processing the audio. This transitions into Data Augmentation, boosting our data variety, followed by Dataset Preparation to organize our learning phases. The Neural Network Architecture Design is followed by our DNN, GRU, and Bi-LSTM setups, flowing into Model Training with Early Stopping for optimized learning. From there, Model Validation and Model Testing ovals guide us through performance checks, culminating in Visualization (Loss Curves, Confusion Matrix) to show a clear picture of success and errors. The downward arrows

connecting each step highlight the logical progression, ensuring every phase builds on the last, creating a cohesive and thorough approach to developing our voice command recognition system.

1. Data Acquisition and Preprocessing

The study utilized the balanced Google Speech Commands Dataset with six command classes which is ”right”, ”off”, ”left”, ”on”, ”down”, and ”up”. MFCCs were extracted as primary audio features, involving pre-emphasis filtering, framing with a Hamming window, Fast Fourier Transform (FFT), Mel filter bank processing, and Discrete Cosine Transform (DCT). To improve generalization, data augmentation techniques like pitch shifting, time stretching, and additive noise injection were applied. The dataset was stratified into training (70%), validation (15%), and testing (15%) subsets. MFCC matrices were standardized using zero-padding and Z-score normalization, and first-order and second-order derivatives of MFCCs were computed to capture dynamic speech properties.

2. Model Architectures

Three deep learning architectures were evaluated: DNN, GRU, and Bi-LSTM models. LSTM networks were chosen to address the vanishing gradient problem common in Recurrent Neural Networks (RNNs). Each LSTM unit employs forget, input, and output gates to control information flow. The Bi-LSTM configuration was specifically used to capture both past and future temporal contexts by combining forward and backwards passes.

3. Optimization and Evaluation

Model parameters were optimized using the Adam optimizer to effectively capture patterns in sequential audio data. The SoftMax activation function was used in the output layer for class probabilities [15]and training minimized the categorical cross-entropy loss function. Performance was evaluated through training and validation accuracy, loss curves, confusion matrices, and supplementary metrics such as precision, recall, and F1-score.

III. RESULTS AND DISCUSSIONS

A. Results

The DNN model begins with a Flatten layer that converts the input shape of (None, 63, 60) into a 3,780-dimensional feature vector for processing MFCC-extracted audio data. It includes a Dense layer that reduces the feature space to 256 units with a 25% dropout to mitigate overfitting. Additional Dense and dropout layers compress the representation, concluding with a 6-unit softmax output layer to classify six commands.

In contrast, the GRU model aims to exploit temporal dependencies. It starts with a GRU layer producing an output shape of (None, 63, 128), followed by dropout and a second GRU layer, which condenses to (None, 64). After more dropout and a Dense layer with 32 units, it also ends with a 6-unit softmax output layer.

Finally, the Bi-LSTM architecture employs a bidirectional layer, capturing full sequence context, followed by regularization, an LSTM layer condensing to (None, 64), and further dropout before a final Dense layer and softmax output. Each model is optimized using the Adam optimizer with an ExponentialDecay learning rate schedule.

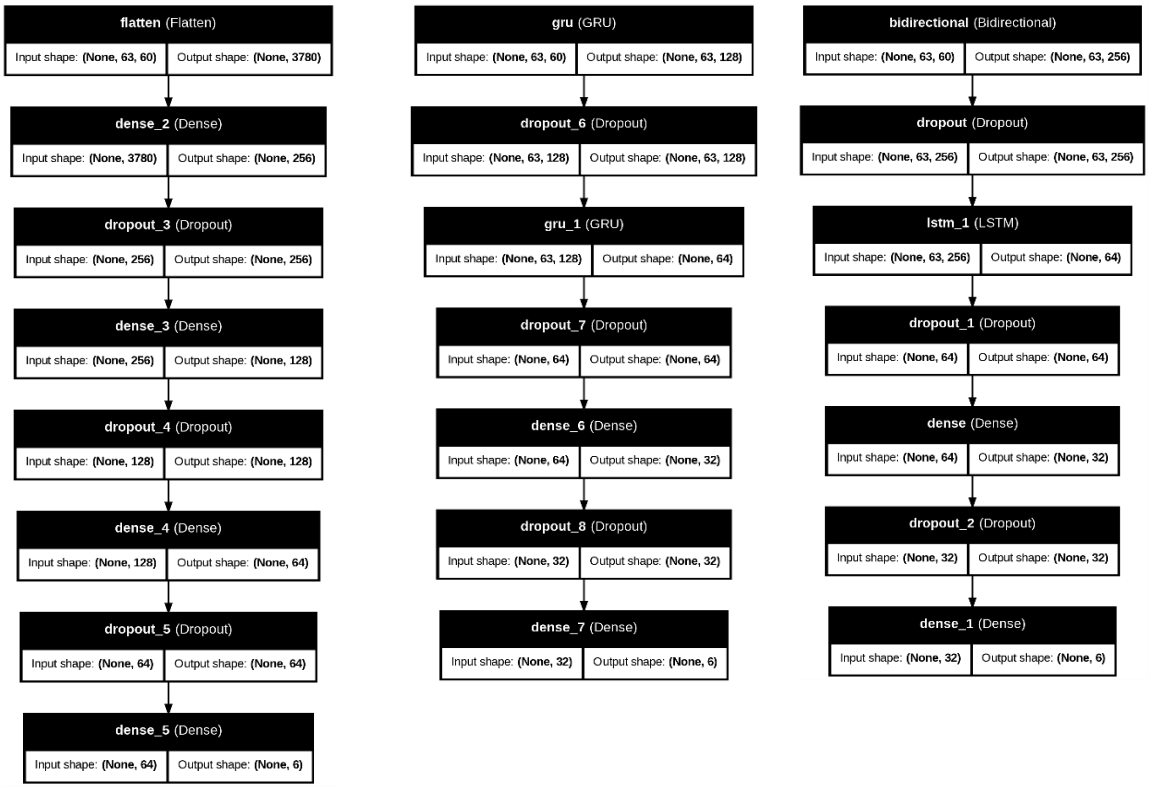

Figure 2. DNN, GRU, and Bi-LSTM Model Architecture

Training results were evaluated over 50 epochs, with performance metrics visualized through loss and accuracy evolution plots. The training loss decreased from an initial value of 1.6 to approximately 0.4, indicating robust learning as the model adjusted its weights to fit the data. The validation loss, starting at 1.2, stabilized around 0.6 with minor fluctuations, suggesting convergence with some variance likely due to the dataset's diversity. Accuracy metrics showed the training accuracy rising from 0.4 to 0.95, reflecting a strong fit to the training set, while the validation accuracy increased from 0.5 to 0.85, plateauing with slight dips. This plateau, managed by early stopping, indicates a well-generalized model, though the validation variance suggests potential sensitivity to data variations that warrant further investigation.

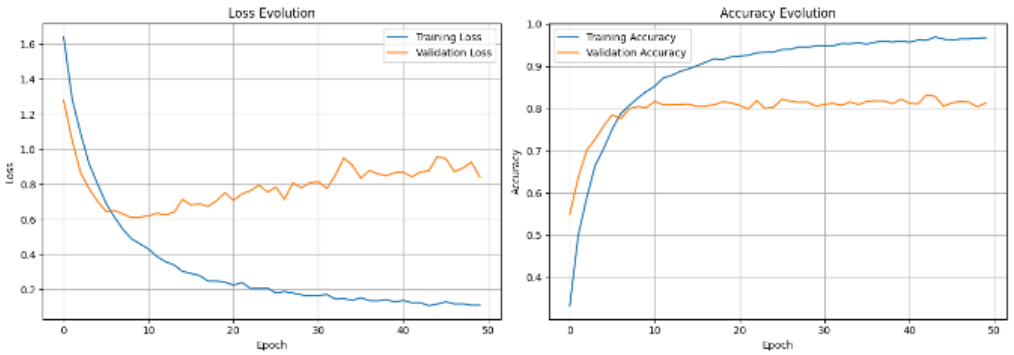

Figure 3. Training and Validation for DNN Model

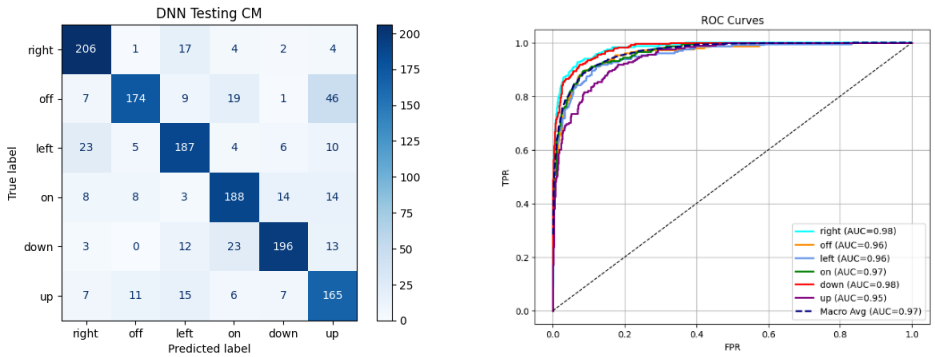

Figure 4. Testing the Confusion Matrix and ROC-AUC Curve DNN Model

Testing on a balanced dataset of 1,418 samples showed the model's effective classification abilities, with an overall accuracy of 79%. The confusion matrix indicated strong performance for classes like "right", "on", and "down", while "off" and "up" faced challenges, notably misclassifying "off" samples as "up". The F1-scores ranged from 0.71 for "up" to 0.84 for "right", and the ROC curves reflected high AUC values, confirming discriminative power despite lower performance on certain classes. Training results over 50 epochs revealed a decrease in loss and a rise in accuracy, indicating robust learning. However, signs of overfitting were present as validation loss increased after a point, suggesting the need for continued refinement in feature extraction and model architecture.

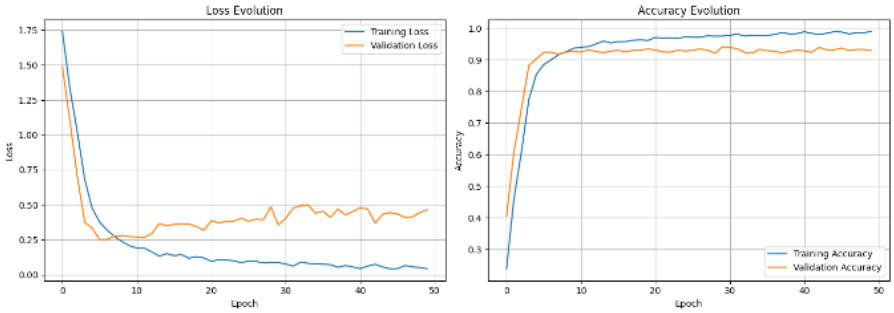

Figure 5. Training and Validation for GRU Model

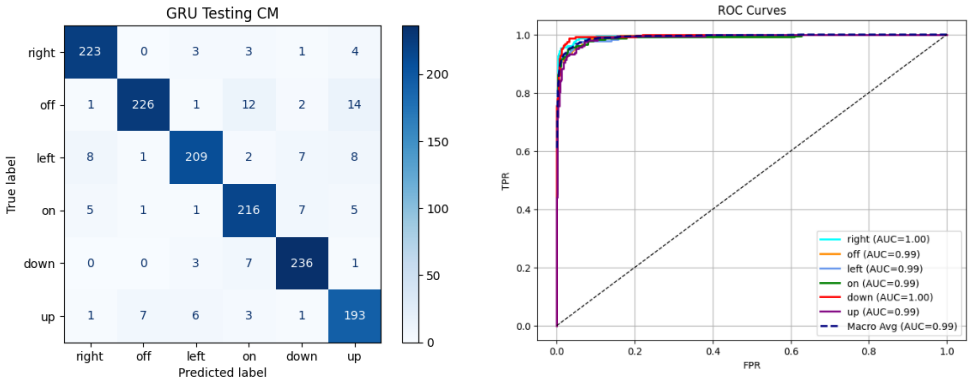

Figure 6. Testing the Confusion Matrix and ROC-AUC Curve GRU Model

Testing on a balanced dataset of 1,418 samples indicated strong model performance, with an overall accuracy of 92% and impressive class separation, as shown in the confusion matrix. True positives included 223 for "right", 226 for "off", 216 for "on", and 236 for "down", with minor misclassifications suggesting some feature overlaps. Precision ranged from 0.86 for "up" to 0.96 for "off", while recall varied between 0.88 for "off" and 0.96 for "down". ROC curve AUC values were excellent, averaging 0.99. Training over 50 epochs showed a decline in training loss from 4.0

to 0.5 and an increase in training accuracy from 0.3 to 0.95, while validation loss stabilized around 0.6 and accuracy reached 0.9, suggesting effective learning and good generalization, despite some validation variance that may need further examination.

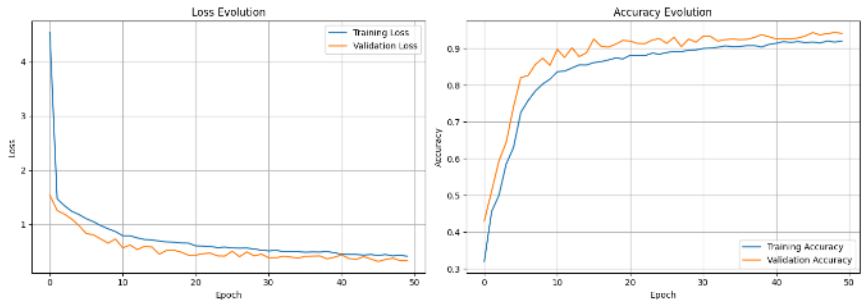

Figure 7. Training and Validation of the Proposed LSTM-Based Model

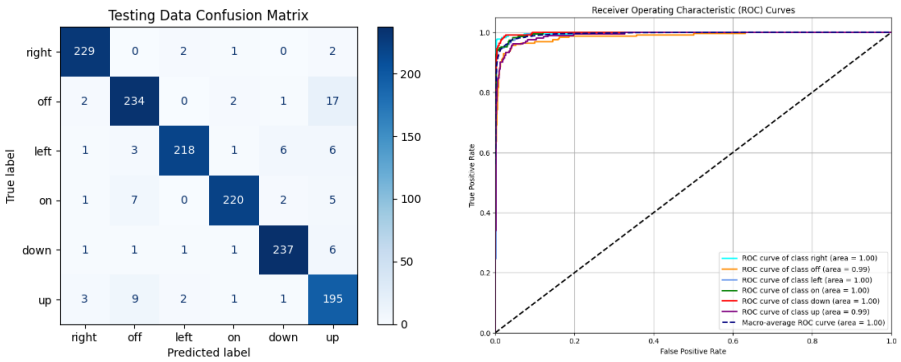

Figure 8. Testing the Confusion Matrix and ROC-AUC Curve of the Proposed LSTM-Based Model

Testing on a balanced dataset of 1,418 samples showed the model's exceptional performance. The confusion matrix indicated strong class separation, with "right" achieving 229 true positives out of 234, "off" 234 out of 256, "on" 220 out of 235, and "down" 237 out of 247. Some misclassifications were noted, such as 17 "off" as "up" and 7 "on" as "off", suggesting minor feature overlaps. The overall accuracy was 94%, with precision from 0.84 (up) to 0.98 (left) and recall from 0.91 (off) to 0.98 (right). F1-scores ranged from 0.88 (up) to 0.97 (right), with an average of 0.94 across classes. The ROC curves supported these findings, with AUC values close to 1.00 for most classes, confirming the model's strong discriminative power while indicating potential improvements for the observed misclassifications.

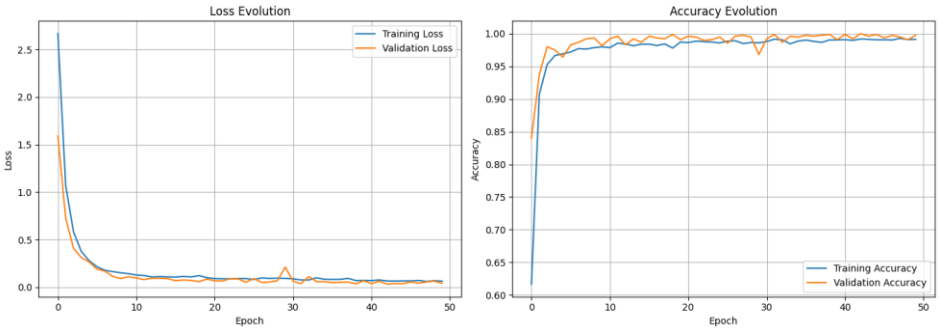

Figure 9. Training and Validation of the Proposed LSTM-Based Model for User Identification

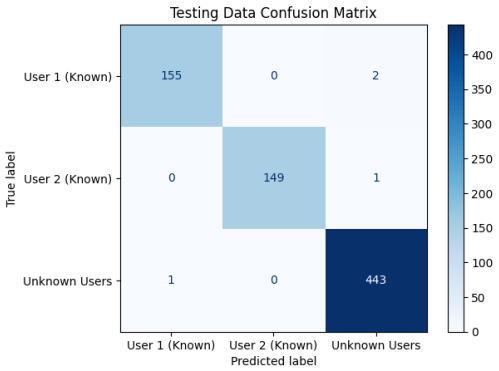

Figure 10. Confusion Matrix of the Proposed LSTM-Based Model for User Identification

The voice command system was enhanced to prioritize speaker identification for "User 1", "User 2", and "Unknown User" using a Bi-LSTM architecture with Layer Normalization. This approach significantly improved training stability and convergence, even with imbalanced class distribution, achieving a drop in training and validation losses from around 3.4 to below 0.1, while accuracy reached 99.87%. The model's computational efficiency also improved, reducing training time for 50 epochs to about 1 minute and 40 seconds, compared to 28 minutes for the original LSTM model, demonstrating its effectiveness for real-time speaker identification in smart lighting systems.

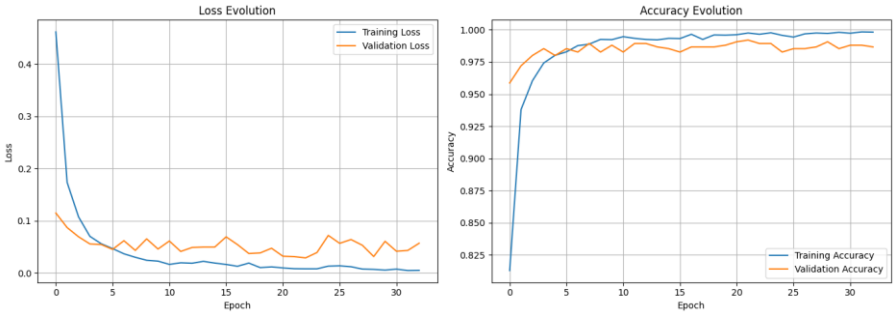

Figure 11. Training and Validation of the DNN Model for User Identification

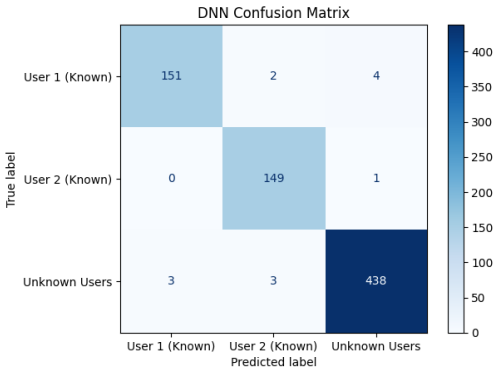

Figure 12. Confusion Matrix of the DNN Model for User Identification

In the DNN model for speaker identification, training accuracy improved from 70% to nearly 100%, with validation accuracy peaking at about 99%. Loss curves showed effective generalization, decreasing from 0.75 to below 0.04, indicating minimal overfitting. The confusion matrix highlighted strong performance, with precision and recall for "User 1" and "User 2" around 97–99%, and an overall accuracy of 98.27%. The model trained efficiently, taking approximately 1 minute and 40 seconds over 50 epochs, making it suitable for real-world applications in smart lighting systems with reliable speaker identification.

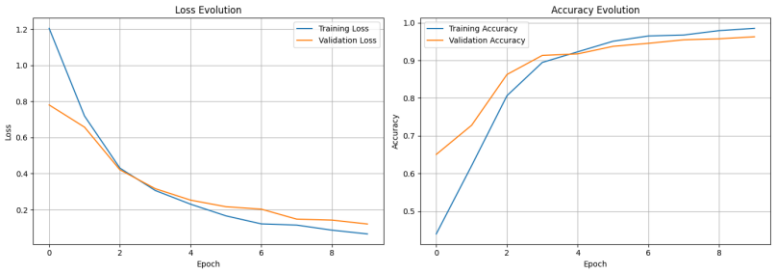

Figure 13. Training and Validation of the GRU Model for User Identification

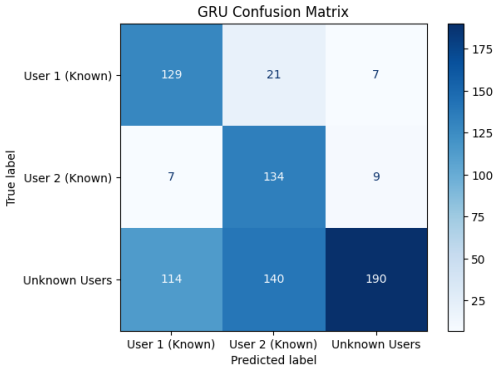

Figure 14. Confusion Matrix of the GRU Model for User Identification

The GRU model for speaker identification showed promising training (98%) and validation (96%) accuracy initially, but rising validation loss indicated overfitting, likely due to class imbalance and speaker variability. It struggled particularly with the "Unknown Users" class, resulting in low macro precision (0.63), recall (0.71), F1-score (0.61), and 60% overall accuracy. Despite training efficiency comparable to DNN, GRU underperformed relative to DNN and Bi-LSTM, highlighting the need for architectural or regularization improvements.

B. Discussions

This study evaluates three models for voice command recognition using the Google Speech Commands Dataset, which includes commands like "right", "off", "left", "on", "down", and "up". The models DNN, GRU, and Bi-LSTM demonstrate progressive improvements in handling sequential data. The DNN, a feedforward model using Dense layers on a 3,780-dimensional input, achieved 79% accuracy but showed significant overfitting. GRU improved accuracy to 92% but also overfit, reaching 100% training accuracy and 90% validation. The Bi-LSTM achieved the highest accuracy at 94%, with better generalization (95% training, 90% validation accuracy), owing to its ability to capture bidirectional temporal dependencies. Computational demands varied: the DNN had the lowest cost, with complexity approximately $O(n \cdot d^2)$, where n is the sample count and d the maximum neuron count per layer. These results reflect a trade-off between computational efficiency and performance, with Bi-LSTM offering the best generalization and accuracy at a higher training cost.

TABLE 1. RESULTS VOICE COMMAND RECOGNITION SUMMARY

Model	Test Accuracy	Overfitting Signs	Computational Complexity	Average Training Time
DNN	79%	Yes	$O(n \cdot d^2)$	3.06 sec/epochs
GRU	92%	Yes	$O(n \cdot s \cdot u^2)$	28.64 sec/epochs
Proposed LSTM-Based	94%	No	$O(2n \cdot s \cdot u^2)$	34.52 sec/epochs

TABLE 2. RESULTS VOICE CHARACTERISTIC IDENTIFICATION SUMMARY

Model	Test Accuracy	Overfitting Signs	Computational Complexity	Average Training Time
DNN	98%	Yes	$O(n \cdot d^2)$	~3 sec/epochs
GRU	60%	Yes	$O(n \cdot s \cdot u^2)$	~25 sec/epochs
Proposed LSTM-Based	99%	No	$O(2n \cdot s \cdot u^2)$	~30 sec/epochs

The Bi-LSTM model outperformed both DNN and GRU baselines with a test accuracy of 94% and showed no signs of overfitting (Table 1). In contrast, the DNN model, while training faster (~2.5 minutes), achieved only 79% accuracy and exhibited clear overfitting. The GRU model reached 92% accuracy but began to overfit after epoch 12 and required ~24 minutes of training. Although Bi-LSTM took the longest to train (~29 minutes), it maintained stable validation performance, indicating superior generalization. A paired t-test confirmed that both GRU and Bi-LSTM significantly outperformed DNN ($p < 0.001$), while no significant difference was found between Bi-LSTM and GRU ($p = 0.9192$), highlighting Bi-LSTM's practical robustness. For speaker identification (Table 2), Bi-LSTM achieved the highest macro F1-score (0.99), followed by DNN (0.98) and GRU (0.61). Paired t-tests showed both Bi-LSTM and DNN significantly outperformed GRU ($p < 0.01$), and Bi-LSTM significantly outperformed DNN ($p \approx 0.0$), confirming its superior generalization across user classes. The command "up" consistently had the lowest classification accuracy, likely due to its short duration and phonetic similarity to "off" and "on," increasing spectral overlap. Compared to previous Bi-LSTM-based speech models, Kumar and Aziz [16] reported 90% accuracy on a speech command task, while Pandiammal et al. [17] achieved 91.45% in emotion classification using MFCC features. These comparisons suggest that the proposed model's performance is within the upper range of prior results, while also offering improvements in speaker generalization and robustness.

IV. CONCLUSION

This study demonstrated that a Bi-LSTM model, combined with MFCC-based data augmentation, can effectively classify voice commands for smart lighting systems, achieving 94% accuracy and outperforming baseline DNN and GRU models. The results support the hypothesis that bidirectional temporal modeling enhances generalization in variable acoustic conditions. Theoretically, the findings confirm the suitability of recurrent models for sequential audio tasks, while practically, they indicate the model's potential for robust voice interfaces in IoT applications. Nevertheless, the experiments were conducted on balanced datasets in controlled environments, which may not fully reflect real-world complexity. Future research should explore the integration of attention mechanisms, test the system in noisy or embedded settings, and investigate multimodal extensions combining voice with other input types. Such efforts would address current limitations and advance the deployment of reliable, context-aware voice-controlled systems.

REFERENCE

[1] S. R. Swamy, K. S. Nandini Prasad, and P. Tripathi, "Smart Home Lighting System," in *Proceedings of the 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing, ICSIDEMPC 2020*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 75–81. doi: 10.1109/ICSIDEMPC49020.2020.9299585.

[2] P. Kumar, P. Rai, and D. H. B. Yadav, "Smart lighting and switching using Internet of Things," in *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data*

Science and Engineering, Institute of Electrical and Electronics Engineers Inc., Jan. 2021, pp. 536–539. doi: 10.1109/Confluence51648.2021.9377078.

[3] M. Huh, R. Ray, and C. Karnei, "A Comparison of Speech Data Augmentation Methods Using S3PRL Toolkit," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2303.00510

[4] Tedyyana, Agus, Osman Ghazali, and Onno W. Purbo. "Machine learning for network defense: automated DDoS detection with telegram notification integration." *Indonesian Journal of Electrical Engineering and Computer Science* 34.2 (2024): 1102.

[5] I. Malashin, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin, "Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review," 2024, doi: 10.3390/polym.

[6] Z. Li, A. Basit, A. Daraz, and A. Jan, "Deep causal speech enhancement and recognition using efficient long-short term memory Recurrent Neural Network," *PLoS One*, vol. 19, no. 1 January, Jan. 2024, doi: 10.1371/journal.pone.0291240.

[7] N. M. Rezk, M. Purnaprajna, T. Nordstrom, and Z. Ul-Abdin, "Recurrent Neural Networks: An Embedded Computing Perspective," 2020, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2020.2982416.

[8] A. Alex, L. Wang, P. Gastaldo, and A. Cavallaro, "Data augmentation for speech separation," *Speech Commun*, vol. 152, Jul. 2023, doi: 10.1016/j.specom.2023.05.009.

[9] J. Galić, B. Marković, Đ. Grozdić, B. Popović, and S. Šajić, "Whispered Speech Recognition Based on Audio Data Augmentation and Inverse Filtering," *Applied Sciences (Switzerland)*, vol. 14, no. 18, Sep. 2024, doi: 10.3390/app14188223.

[10] A. K. Biswal, D. Singh, and B. K. Pattanayak, "IoT-Based Voice-Controlled Energy-Efficient Intelligent Traffic and Street Light Monitoring System," in *Lecture Notes in Networks and Systems*, vol. 151, Springer Science and Business Media Deutschland GmbH, 2021, pp. 43–54. doi: 10.1007/978-981-15-8218-9_4.

[11] S. Nosouhian, F. Nosouhian, and A. Kazemi Khoshouei, "A Review of Recurrent Neural Network Architecture for Sequence Learning: Comparison between LSTM and GRU," Jul. 12, 2021. doi: 10.20944/preprints202107.0252.v1.

[12] K. E. ArunKumar, D. V. Kalaga, C. Mohan Sai Kumar, M. Kawaji, and T. M. Brenza, "Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells, autoregressive Integrated moving average (ARIMA), seasonal autoregressive Integrated moving average (SARIMA) for forecasting COVID-19 trends," *Alexandria Engineering Journal*, vol. 61, no. 10, pp. 7585–7603, Oct. 2022, doi: 10.1016/j.aej.2022.01.011.

[13] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2022.3223444.

[14] D. Prabakaran and S. Sriuppili, "Speech processing: MFCC based feature extraction techniques - An investigation," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1742-6596/1717/1/012009.

[15] Z. Xie, X. Wang, H. Zhang, I. Sato, and M. Sugiyama, "Adaptive Inertia: Disentangling the Effects of Adaptive Learning Rate and Momentum," Jun. 2022, [Online]. Available: http://arxiv.org/abs/2006.15815

[16] D. Kumar and S. Aziz, "Performance Evaluation of Recurrent Neural Networks-LSTM and GRU for Automatic Speech Recognition," in *2023 International Conference on Computer, Electronics and Electrical Engineering and their Applications, IC2E3 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/IC2E357697.2023.10262561.

[17] K. Sankara Pandiammal, S. Karishma, K. Harine Sakthe, V. Manimaran, S. Kalaiselvi, and V. Anitha, "Emotion Recognition from Speech - an LSTM approach with the Tess Dataset," in *2024 5th International Conference on Innovative Trends in Information Technology, ICITIIT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICITIIT61487.2024.10580351.